INFORMATION-ALIGNED PREFERENCE OPTIMIZATION FOR DIRECT PREFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has become a standard approach for aligning large language models with human feedback. Recent work has addressed its sensitivity to response length, yet equalizing token counts alone leaves open which tokens should contribute to preference learning. In this work, we study this question through an information-theoretic lens and introduce Information-Aligned Preference Optimization (IAPO). IAPO uses model entropy as a practical ranking signal, masking low-entropy positions in the longer response to match the retained token count of the shorter response. Our analysis shows that true conditional entropy upper-bounds token-level preference information and that top-K selection minimizes a discarded-entropy surrogate under a fixed token budget. Experiments across multiple model families and scales show that IAPO improves over DPO and achieves strong average performance against recent variants. Matched-budget ablations further demonstrate the benefit of entropy-guided selection beyond length equalization, supporting selective token aggregation as an effective strategy for preference optimization.

1 INTRODUCTION

Large language models (LLMs) (Scao et al., 2022; Achiam et al., 2023; Touvron et al., 2023) are commonly aligned through preference learning (Christiano et al., 2017; Ziegler et al., 2019; Askell et al., 2021; Ouyang et al., 2022; Chung et al., 2024; Stiennon et al., 2020). Reinforcement Learning from Human Feedback (RLHF) typically combines reward modeling with online policy optimization (Ouyang et al., 2022; Dong et al., 2024), introducing computational cost and potential reward-model exploitation (Gao et al., 2023; Weng, 2024). Direct Preference Optimization (DPO) (Rafailov et al., 2023) instead fits pairwise preferences through policy-to-reference log-likelihood ratios, without a separate reward model. However, DPO can exhibit length bias: improvements in judged preference may partly reflect increased verbosity (Shen et al., 2023; Singhal et al., 2023; Park et al., 2024).

DPO sums token-level log-likelihood ratios across each response. This raises two questions: how should the number of terms be balanced between responses of different lengths, and which positions should enter the comparison? Existing approaches include explicit length regularization in R-DPO (Park et al., 2024), average log-probability scoring in SimPO (Meng et al., 2024), and random token down-sampling in SamPO (Lu et al., 2024). SamPO provides a particularly direct control for the first question by equalizing the retained-token budget.

We focus on the second question within this equal-budget setting. Preference evidence can be localized in facts, reasoning steps, or task constraints, motivating selective treatment of token positions (Qi et al., 2025; Samineni et al., 2025). Predictive entropy offers a computable ranking signal: low entropy indicates a predictable continuation, whereas high entropy indicates model uncertainty. A preliminary diagnostic finds an association between entropy and the chosen/rejected log-probability gap, but this diagnostic does not identify a token's causal contribution to preference.

We propose Information-Aligned Preference Optimization (IAPO). For each pair, IAPO keeps all positions in the shorter response and selects the same number of highest-entropy positions from the longer response. Only loss terms are masked; retained-token probabilities still condition on the full response prefix. The entropy scores and masks are recomputed under the current policy during

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training. Compared with SamPO, the central change is replacing uniform random selection with entropy ranking at the same token budget.

Our analysis states the limits of this motivation explicitly. Under the true data distribution, conditional entropy upper-bounds a token's conditional mutual information with the preference label. This one-sided bound excludes large information at low-entropy positions, but does not guarantee preference relevance at high-entropy positions. We also establish a discarded-information bound for a fixed subset and show that top-K selection minimizes a discarded-entropy surrogate. These results motivate the selection rule; they do not establish an information-preservation or convergence guarantee for adaptive masking with model entropy.

We evaluate IAPO on open-ended alignment, general understanding, and mathematical reasoning tasks. The most direct evidence comes from matched-budget ablations: on Qwen2.5-Math-7B, full-sequence DPO, random-K, bottom-K, and IAPO achieve average scores of 49.4, 51.5, 41.2, and 54.1, respectively. Three independent training runs further support the DPO/random-K/IAPO ordering in this setting. Individual benchmark winners vary, so our empirical claim concerns average performance and the controlled selection comparison.

Our contributions are:

- An equal-budget preference objective that ranks token positions by current-policy entropy while preserving full prefixes.
- An analysis distinguishing true-distribution information bounds from optimization of a computable entropy surrogate.
- Multi-model evaluation and matched-budget ablations, including independent training repeats for the central mathematical-reasoning comparison.

2 Preliminaries and Background

We introduce the notation and background needed for the development of IAPO. In particular, we set up pairwise preference learning, review DPO and representative approaches to length bias, and introduce the entropy quantities that will be used in the method and theory sections.

2.1 Preliminaries

We consider preference learning from paired responses. Let $\mathcal{D} = \{(x_{(i)}, y^w_{(i)}, y^l_{(i)})\}_{i=1}^N$ denote the pairwise-preference ranked dataset, where $x_{(i)}$ represents a prompt, and $y^w_{(i)}$ and $y^l_{(i)}$ correspond to the chosen and rejected responses, respectively. When discussing response length later, we will use y^{long} and y^{short} for the longer and shorter responses in a pair, independent of whether they are chosen or rejected. A policy $\pi_\theta(y|x)$ defines a distribution over responses, and $\pi_{\text{ref}}(y|x)$ denotes a fixed reference policy, typically obtained via supervised fine-tuning. DPO optimizes the policy by contrasting the likelihood ratios of preferred and dispreferred responses under the policy and reference models. The DPO objective is given by

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathop{\mathbb{E}}_{(x,y^w,y^l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y^w|x)}{\pi_{\text{ref}}(y^w|x)} - \beta \log \frac{\pi_\theta(y^l|x)}{\pi_{\text{ref}}(y^l|x)} \right) \right], \tag{1}$$

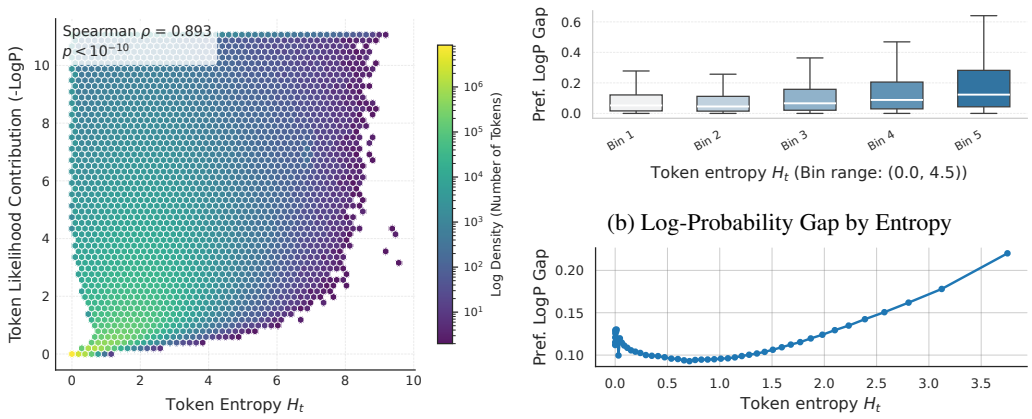
where β controls the strength of preference supervision.

2.2 Length Regularization in DPO

Although DPO is formulated at the sequence level, expanding the sequence likelihood reveals an implicit token-wise aggregation. Specifically, the log-likelihood ratio in the DPO objective (1) can be decomposed as

$$\log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} = \sum_{t=1}^{|y|} \log \frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\text{ref}}(y_t|x, y_{<t})}. \tag{2}$$

This decomposition assigns equal weight to all tokens in a response. Responses of different lengths therefore contribute different numbers of terms, motivating explicit treatment of length in preference



(a) Negative Log-Probability vs Entropy

(b) Log-Probability Gap by Entropy

(c) Mean Log-Probability Gap vs Entropy

Figure 1: **Token-level diagnostics at the Llama3-8B SFT checkpoint.** (a) Observed-token negative log-probability versus predictive entropy. (b) Absolute chosen/rejected log-probability gaps grouped by rejected-side predictive entropy. (c) Mean gaps across entropy bins. The trend is broadly increasing in the moderate-to-high entropy regime, with a low-entropy shoulder. The gap is a likelihood diagnostic, not a token-level preference attribution.

optimization. Several recent works (Park et al., 2024; Lu et al., 2024; Meng et al., 2024) have proposed corrections to mitigate this effect, including explicit length regularization and token down-sampling. This formulation makes clear that preference optimization depends on token-level log-likelihood contributions, which motivates examining how information is distributed across tokens.

2.3 Entropy Mechanism in LLMs

Entropy in information theory was introduced by Shannon (1948), which quantifies the expected uncertainty or information associated with a random variable's possible outcomes. In the context of LLMs, token-level entropy is widely used to characterize predictive uncertainty and information density during generation (Cui et al., 2023; Wang et al., 2025). For a response y and prompt x, the token-level entropy under the policy π at position t is defined as

$$H_t = -\sum_{v \in \mathcal{V}} \pi(v|x, y_{<t}) \log \pi(v|x, y_{<t}), \tag{3}$$

where $\mathcal{V}$ denotes the vocabulary. Predictive entropy measures model uncertainty: low entropy indicates a concentrated next-token distribution, while high entropy indicates a more dispersed distribution. Its usefulness for ranking preference-learning positions is an empirical question.

3 Method

We examine token-level diagnostics and then define an equal-budget preference objective using entropy to rank positions.

3.1 Token-Level Entropy Diagnostics

The statistics in Figure 1 use an SFT-trained Llama3-8B model (AI@Meta, 2024) on UltraFeedback (Cui et al., 2023), before preference optimization. Panel (a) relates predictive entropy to observed-token negative log-probability, with Spearman correlation $\rho = 0.893$ ($p < 10^{-10}$). This is a likelihood diagnostic, not a measure of preference relevance.

Panels (b)–(c) compare entropy with the absolute chosen/rejected log-probability gap at corresponding response positions, $|\log \pi_\theta(y_t^w \mid x, y_{<t}^w) - \log \pi_\theta(y_t^l \mid x, y_{<t}^l)|$, using rejected-side predictive entropy. The binned gap generally increases with entropy, but the low-entropy shoulder prevents a globally

monotonic interpretation. The two probabilities condition on different response histories, so their gap is an operational diagnostic rather than token-level mutual information or causal preference attribution. These observations motivate testing entropy ranking empirically; they do not establish its validity throughout adaptive training. Complementary token visualizations appear in Appendix B.

3.2 Information-Aligned Preference Optimization

For a pair (x, y^w, y^l), set $K = \min(|y^w|, |y^l|)$. For each response y, compute current-policy entropy H_t using Equation (3) and define

$$m_t^y = \mathbf{1}\{t \in \mathrm{Top}_K(H_1, \ldots, H_{|y|})\}. \tag{4}$$

Thus, every position in the shorter response is retained, and exactly K positions are selected from the longer response. Equal-length pairs retain both responses in full. The budget counts valid response positions, excluding prompt and padding positions.

Define the masked log-likelihood ratio

$$A_\theta(x, y; m^y) = \sum_{t=1}^{|y|} m_t^y \log \frac{\pi_\theta(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})}. \tag{5}$$

The same response-specific mask is applied to policy and reference terms. The complete pairwise objective is

$$\mathcal{L}_{IAPO}(\theta) = -\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}}\left[\log \sigma\left(\beta\left[A_\theta(x, y^w; m^w) - A_\theta(x, y^l; m^l)\right]\right)\right]. \tag{6}$$

Entropy scores and discrete masks are recomputed from the current policy at each update, and masks are held fixed for backpropagation. Masking removes loss terms only: all retained-token probabilities use the full original prefix, including tokens whose loss terms are discarded. The rule introduces no selection-budget hyperparameter beyond the shorter-response length.

4 Information Bounds and an Entropy Surrogate

We distinguish information bounds under the true data distribution from optimization of an entropy surrogate. The latter motivates the implemented selection rule without providing an adaptive information-preservation guarantee.

4.1 Setup

For the information bounds, fix a response length T and let P denote the true data distribution over (X, Y, Z), where X is a prompt, $Y = (Y_1, \ldots, Y_T)$ is a response, and $Z \in \{+, -\}$ indicates whether Y is preferred or dispreferred in a pairwise comparison. For any subset $S \subseteq [T]$, let $Y_S := \{(t, Y_t) : t \in S\}$ denote the retained token-position pairs. We write the *true local entropy* as $h_t^\star(x, y_{<t}) := H_P(Y_t \mid X = x, Y_{<t} = y_{<t})$ and the *policy entropy* as $\widehat{h}_t(x, y_{<t}) := H_{\pi_\theta}(Y_t \mid X = x, Y_{<t} = y_{<t})$. For a response y and token budget $K \le |y|$, IAPO selects

$$\widehat{S}_K(x, y) \in \arg\max_{\substack{S\subseteq[|y|]\\|S|=K}} \sum_{t\in S} \widehat{h}_t(x, y_{<t}).$$

Equivalently, IAPO discards the $|y| - K$ positions with the smallest policy entropy.

4.2 Entropy as an Exclusion Bound

The first result gives a *one-sided* information-theoretic property of entropy.

Proposition 4.1 (Token-level exclusion bound)**.** *For every token position t,*

$$I_P(Y_t; Z \mid X, Y_{<t}) \;\le\; H_P(Y_t \mid X, Y_{<t}).$$

Equivalently, for every fixed prefix $(x, y_{<t})$*,*

$$I_P(Y_t; Z \mid X = x, Y_{<t} = y_{<t}) \;\le\; h_t^\star(x, y_{<t}).$$

Proposition 4.1 is one-sided: small true conditional entropy limits preference information at a position. High entropy alone gives no lower bound on that information, and policy entropy need not equal true entropy.

4.3 From Token-Level to Subsequence-Level Discarded Information

For masking, the key quantity is the residual information in the *entire discarded subsequence*, not in any single token. The next proposition lifts the token-level bound to this subsequence-level quantity.

Proposition 4.2 (Discarded-subsequence bound)**.** *For any deterministic subset $S \subseteq [T]$,*

$$I_P(Y_{S^c}; Z \mid X, Y_S) \;\le\; \sum_{t \in S^c} H_P(Y_t \mid X, Y_{<t}).$$

For a fixed S, the right-hand side averages local entropies over P; it is not a bound for each realized response. The associated expected posterior-closeness result is given in Corollary A.1. An adaptive subset $S(X, Y, \theta)$ is outside the scope of this proposition.

4.4 Fixed-Budget Optimization of an Entropy Surrogate

Motivated by Proposition 4.2, we define a sample-level entropy objective. For a fixed prompt-response pair (x, y), define the true-entropy surrogate

$$U_x^\star(y, S) := \sum_{t \notin S} h_t^\star(x, y_{<t}), \qquad S \subseteq [|y|].$$

This is a pathwise surrogate formed from local entropies; we do not claim that it upper-bounds the preference information discarded by a sample-dependent mask.

Proposition 4.3 (Fixed-budget optimality of top-K under the entropy surrogate)**.** *Fix a prompt-response pair (x, y) and a token budget $K \le |y|$. If*

$$S_K^\star(x, y) \in \arg\max_{\substack{S \subseteq [|y|] \\ |S| = K}} \sum_{t \in S} h_t^\star(x, y_{<t}),$$

then

$$S_K^\star(x, y) \in \arg\min_{\substack{S \subseteq [|y|] \\ |S| = K}} U_x^\star(y, S).$$

Thus, top-K minimizes the sum of discarded true local entropies under a fixed budget.

This is an optimization result for the stated surrogate, not for mutual information itself. Setting $K = \min(|y^w|, |y^l|)$ gives the equal-budget pairwise rule; see Corollary A.2.

4.5 Practical Proxy and Scope of the Claim

In practice, IAPO substitutes current-policy entropy $\widehat{h}_t$ for $h_t^\star$. Under the assumption $|\widehat{h}_t - h_t^\star| \le \delta$ at every response position, Proposition A.3 bounds the selected subset's true-entropy surrogate gap by $2K\delta$. This is a conditional result; our experiments do not estimate or verify a uniform bound on δ.

Remark 4.4 (Scope). The fixed-subset information bound, surrogate optimality, and conditional proxy bound have distinct scopes. They do not guarantee that high-entropy tokens are preference-relevant, characterize the optimal policy under adaptive masking, or establish convergence of the implemented updates. Moreover, loss masking preserves full prefixes, whereas the fixed-subset information bound concerns observing a subsequence.

5 Related Work

Preference optimization. RLHF combines preference-based reward modeling with policy optimization (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). DPO (Rafailov et al., 2023) uses a direct pairwise loss, while ORPO (Hong et al., 2024) uses a reference-free odds-ratio objective and iterative DPO (Xiong et al., 2023; Guo et al., 2024) refreshes preference data during training. Length-sensitive scoring motivates R-DPO's explicit length term (Park et al., 2024) and SimPO's average log-probability reward with a target margin (Meng et al., 2024). SamPO (Lu et al., 2024) randomly samples positions to match the retained-token count; IAPO changes this sampling rule to entropy ranking.

Table 1: **Results on open-ended benchmarks.** We report performance on Arena-Hard v0.1, and AlpacaEval 2.0. WR denotes win rate, LC denotes the length-controlled win rate, and the Arena-Hard 95% CI column reports the lower and upper deviations around the WR estimate. We **bold** the best scores and underline the second best distinct scores; ties at the displayed precision share the same marking.

Model	Method	Arena-Hard			AlpacaEval2			
		WR(%)	95% CI	Avg. #tokens	WR(%)	LC(%)	Std Error	Avg. #tokens
Llama3-8B-Base	SFT	3.3	$(-0.6, 0.8)$	389	3.7	5.7	0.66	1000
	DPO	15.9	$(-1.6, 1.6)$	542	13.3	15.5	1.20	1708
	SimPO	23.4	$(-1.9, 1.7)$	704	**18.1**	19.4	1.36	1883
	SamPO	25.4	$(-2.0, 1.1)$	422	14.25	18.74	1.11	1498
	R-DPO	19.0	$(-1.8, 1.6)$	285	3.75	7.35	0.58	608
	IAPO	**27.3**	$(-1.8, 1.5)$	332	14.42	**20.18**	1.13	1309
Llama3-8B-Instruct	Baseline	20.6	$(-2.0, 1.9)$	585	28.4	28.5	1.59	1955
	DPO	26.8	$(-2.2, 2.1)$	522	41.7	43.7	1.74	1890
	SimPO	28.2	$(-1.9, 2.1)$	466	40.4	43.32	1.73	1779
	SamPO	25.2	$(-2.3, 1.8)$	290	28.97	32.01	1.44	1832
	R-DPO	26.6	$(-2.2, 1.8)$	297	28.51	40.82	1.46	1441
	IAPO	**29.9**	$(-1.8, 1.9)$	433	**44.62**	**44.68**	1.33	2010

Table 2: **Results on general benchmarks.** Acc. denotes accuracy. For TruthfulQA, MC1 is single-true answer accuracy and MC2 is the normalized probability mass assigned to true answers. We report IFEval Instruction/Prompt (Loose/Strict Acc.), HellaSwag (Acc.), TruthfulQA (MC1/MC2), and PiQA (Acc.). The last column is the equally weighted arithmetic mean of the eight displayed metrics, rounded to two decimal places. We **bold** the best scores and underline the second best distinct scores; ties at the displayed precision share the same marking.

Model	Method	Instruction		Prompt		HellaSwag	TruthfulQA		PiQA	Average
		Loose(%)	Strict(%)	Loose(%)	Strict(%)	Acc.(%)	MC1(%)	MC2(%)	Acc.(%)	Score(%)
Llama3-8B-Base	SFT	37.17	35.01	23.48	22.18	61.01	31.09	45.31	81.39	42.08
	DPO	43.29	38.85	31.79	26.25	64.26	38.07	53.46	82.10	47.26
	SimPO	**50.96**	**43.88**	**37.71**	**29.76**	61.82	42.47	59.53	81.83	51.00
	SamPO	41.01	39.21	27.73	25.88	63.74	34.39	50.76	81.94	45.58
	R-DPO	46.40	39.81	31.24	24.95	62.57	34.27	48.30	82.10	46.21
	IAPO	48.80	42.81	36.41	**29.76**	**68.31**	**52.14**	**66.03**	**83.57**	**53.48**
Llama3-8B-Instruct	DPO	60.07	51.80	47.13	37.89	60.08	41.86	59.03	**79.82**	54.71
	SimPO	60.67	50.72	45.84	35.86	57.75	43.82	60.68	77.86	54.15
	SamPO	**62.95**	52.04	51.20	38.26	59.57	40.02	57.01	79.43	55.06
	R-DPO	61.95	**52.83**	50.60	**41.70**	55.83	40.88	60.01	**79.82**	55.45
	IAPO	62.23	52.40	**51.39**	40.30	**60.32**	**44.68**	**61.64**	79.05	**56.50**
Gemma2-9B-it	DPO	49.76	47.72	42.14	39.19	53.06	36.84	57.56	71.44	49.71
	SimPO	62.11	59.95	56.56	53.42	49.25	36.35	59.04	69.75	55.80
	SamPO	**73.50**	70.98	61.30	59.89	55.05	35.99	56.68	75.08	61.06
	R-DPO	68.23	65.47	58.23	54.34	53.51	33.90	55.87	73.61	57.90
	IAPO	73.14	**71.10**	**63.96**	**61.37**	**56.15**	**42.35**	**64.22**	**75.19**	**63.44**

Other variants modify different parts of the objective. TDPO (Zeng et al., 2024) controls token-level forward KL divergence, β-DPO (Wu et al., 2024) adapts β and filters examples using batch-level data quality, and BPO (Kim et al., 2025) generalizes the ratio-matching loss through Bregman objectives. Our experiments isolate selection against random sampling at a common budget; they do not establish superiority over these untested methods.

Entropy-based methods. Entropy regularization is widely used in reinforcement learning (Ziebart et al., 2008; Haarnoja et al., 2017; 2018). For LLMs, entropy has been used for unsupervised reasoning fine-tuning (Agarwal et al., 2025), policy-update control (Wen et al., 2024; Cui et al., 2025), and selective token optimization (Wang et al., 2025). IAPO uses current-policy entropy to allocate a fixed token budget within each preference pair.

6 Experiments

We evaluate average benchmark performance across model settings and use matched-budget ablations to test the contribution of entropy ranking beyond length equalization.

6.1 Experimental Setup

Models and preference data. We evaluate Llama3-8B-Base, Llama3-8B-Instruct, and Gemma2-9B-it on textual tasks, and Qwen2.5-Math-1.5B/7B on mathematical reasoning. Text preference data follow the preparation used by SimPO (Meng et al., 2024); the Llama3-8B-Instruct comparison uses `princeton-nlp/llama3-ultrafeedback`. For mathematics, we sample 20K prompts from NuminaMath (LI et al., 2024), generate eight responses per prompt, and rank them with a rule-based reward. The highest- and lowest-reward responses form a preference pair; prompts for which all responses receive the same reward are excluded. All compared methods use the resulting preference pairs within each mathematical model setting.

Baselines and training. We compare with DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), R-DPO (Park et al., 2024), and SamPO (Lu et al., 2024). SamPO supplies the random-K control in Section 6.4, matching the retained-token count of IAPO. Training uses AdamW, cosine learning-rate decay, global batch size 128, and eight NVIDIA A800 GPUs. Textual models are trained for one epoch and mathematical models for two; maximum sequence lengths are 2,048 and 4,096, respectively. Learning rates are 10^{-6} for Llama3 and Qwen2.5-Math and 8×10^{-7} for Gemma2. In the Llama3-8B-Instruct comparison, DPO uses $\beta = 0.01$, while the official SimPO implementation uses $\beta = 2.5$ and $\gamma/\beta = 0.55$. Complete model-level settings and baseline objectives are given in Appendix D.

Evaluation and aggregation. Open-ended evaluation uses Arena-Hard v0.1 (500 prompts) and AlpacaEval 2.0 (805 prompts), with GPT-4 Turbo as judge. We report both raw and length-controlled AlpacaEval win rates, together with generated response lengths. General understanding is evaluated using IFEval, HellaSwag, TruthfulQA, and PiQA. The general benchmark average gives equal weight to the eight displayed metrics, including the four IFEval and two TruthfulQA metrics; it is not an equal-weight average over the four benchmark names. Mathematical evaluation uses zero-shot chain-of-thought prompting and greedy decoding on the eight benchmarks in Table 3, whose accuracies are averaged equally. The independent-training summaries in Section 6.5 are reported separately from the single-run benchmark tables. Prompt templates and evaluator details appear in Appendix D; code is available at `https://anonymous.4open.science/r/IAPO`.

6.2 Results on Textual Understanding Tasks

Tab. 1 reports results on Arena-Hard and AlpacaEval2. For both Llama3-8B settings, IAPO improves over DPO and attains the best Arena-Hard win rate. On AlpacaEval2, IAPO achieves the best length-controlled win rate even though SimPO obtains the best raw win rate on Base, suggesting that the gains are not explained by verbosity alone. Tab. 2 covers a broad mix of conditional benchmarks to assess whether preference optimization gains come without degrading general language capabilities. IAPO achieves the highest average score in each of the three model settings; individual benchmark winners vary. On Gemma2, the average is 63.44 versus 61.06 for SamPO. R-DPO produces shorter responses in the open-ended evaluations, so these results support judged quality rather than a claim of minimizing output length. Open-ended model coverage here is limited to Llama3.

6.3 Results on Mathematical Reasoning Tasks

Table 3 reports average scores of 45.6 and 54.1 for IAPO on Qwen2.5-Math-1.5B and 7B, respectively. The strongest competing displayed averages are 42.8 for SimPO/SamPO at 1.5B and 52.7 for SimPO at 7B. Individual exceptions include SamPO on AMC2023 at 7B and OlympiadBench at 1.5B.

The improvement over DPO is 4.8 and 4.7 percentage points in the displayed averages at 1.5B and 7B, respectively. Against SamPO at 7B, gains are uneven across tasks: Minerva Math increases from 24.3 to 33.5 and AIME2024 from 23.3 to 30.0, while MATH changes from 76.0 to 76.4 and College Math from 45.3 to 45.4. AMC2023 decreases from 62.5 to 60.0. At 1.5B, IAPO leads on six of the eight benchmarks, while SamPO leads on OlympiadBench and College Math; on the latter, SamPO scores 51.3 versus 43.9 for IAPO. The aggregate improvements therefore coexist with task-specific trade-offs, motivating the complete per-benchmark reporting and the independent-training comparison below.

Table 3: **Results on mathematical benchmarks.** We report performance on GSM8K, MATH, Minerva Math, Olympiad Bench, AIME 2024, AMC 2023, GaoKao 2023 En, and College Math. Average is the equally weighted mean of the eight displayed scores, rounded to one decimal place. We **bold** the best scores and underline the second best distinct scores; ties at the displayed precision share the same marking.

Model	Method	GSM8K	MATH	Minerva Math	Olympiad Bench	AIME 2024	AMC 2023	GaoKao 2023 En	College Math	Average
Qwen2.5-Math-1.5B	Baseline	72.1	61.8	15.1	29.2	3.3	40.0	54.3	39.6	39.4
	DPO	73.8	62.2	21.0	30.5	3.3	40.0	54.5	40.9	40.8
	SimPO	74.1	67.2	24.3	30.1	10.0	40.0	54.0	42.6	42.8
	SamPO	72.8	64.6	20.2	**31.0**	6.7	42.5	53.5	**51.3**	42.8
	R-DPO	71.6	62.8	16.5	29.0	10.0	37.5	54.3	38.6	40.0
	IAPO	**75.9**	**67.8**	**26.1**	30.8	**13.3**	**50.0**	**56.6**	43.9	**45.6**
Qwen2.5-Math-7B	Baseline	63.5	65.4	9.2	22.5	23.3	52.5	44.9	31.2	39.1
	DPO	80.0	74.8	21.3	38.1	26.7	52.5	57.1	44.3	49.4
	SimPO	84.8	74.2	29.8	**38.2**	26.7	60.0	63.4	44.4	52.7
	SamPO	84.0	76.0	24.3	36.9	23.3	**62.5**	60.0	45.3	51.5
	R-DPO	72.2	70.6	14.0	32.1	26.7	47.5	54.0	40.4	44.7
	IAPO	**86.4**	**76.4**	**33.5**	37.8	**30.0**	60.0	**63.6**	**45.4**	**54.1**

6.4 Ablation: Token Selection and Masking Design

Table 4: **Ablation on token selection and masking.** Text metrics use Llama3-8B-Base; Math is the eight-benchmark average on Qwen2.5-Math-7B. All scores are percentages.

Method	Arena-Hard	AlpacaEval 2.0		Math
	WR (%)	WR (%)	LC (%)	Average (%)
DPO	15.9	13.30	15.50	49.4
Random-K (SamPO)	25.4	14.25	18.74	51.5
Bottom-K	24.9	7.62	8.66	41.2
Rej-only	25.3	**15.52**	17.61	52.5
Cho-only	26.1	14.25	18.77	53.3
IAPO	**27.3**	14.42	**20.18**	**54.1**

Controlling the retained-token count. Table 4 separates the effects of length equalization, entropy ranking, and one-sided masking. Arena and AE2 report Llama3-8B-Base results, while Math is the Qwen2.5-Math-7B average. R-K denotes SamPO-style uniform random selection, and B-K retains the lowest-entropy positions. For R-K, B-K, and IAPO, each response contributes exactly $K = \min(|y^w|, |y^l|)$ terms, with all positions in the shorter response retained. Random-K improves over full-sequence DPO on every reported metric, including an increase from 49.4 to 51.5 on mathematics. This comparison identifies a benefit from equalizing the budget before introducing entropy ranking.

Controlling which positions are retained. At this common budget, IAPO exceeds Random-K by 1.9 percentage points on Arena-Hard, 1.44 points on AE2 length-controlled win rate, and 2.6 points on the math average. Bottom-K yields 41.2 on mathematics and 8.66 on AE2 length-controlled win rate, below both DPO and Random-K. Thus, the improvement depends on the selection rule as well as the number of retained terms. Because all selected-token probabilities retain the full original prefix, these variants compare loss contributions while preserving the context used to predict each token. The results support entropy ranking in the evaluated settings; they do not establish that every retained high-entropy position carries preference information.

Asymmetric masking. Rej-only applies entropy selection only when the rejected response is longer, while Cho-only applies it only when the chosen response is longer. Both variants outperform DPO on the reported text metrics and on math. Rej-only has the strongest AE2 raw win rate, while Cho-only is stronger on Arena-Hard, AE2 length-controlled win rate, and the math average. Applying the rule to whichever response is longer gives IAPO the best Arena-Hard, length-controlled win rate, and mathematical average in this ablation. Per-benchmark breakdowns in Appendix E also show individual exceptions, including Cho-only on Minerva Math and AMC2023.

6.5 Stability Across Independent Training Runs

We independently train DPO, SamPO-style random-K, and IAPO three times each on Qwen2.5-Math-7B. For each trained model, we first average its accuracies equally over the eight mathematical benchmarks. Table 5 then reports the mean and standard deviation of these three run-level averages for each method. The standard deviations therefore summarize variation across independently trained models; they are not variation across benchmarks or repeated decoding of one checkpoint.

Table 5: **Independent training runs.** Qwen2.5-Math-7B results over three runs per method. Values are mean $\pm$ SD of the eight-benchmark average.

Method	**Average accuracy (%)**
DPO	49.6 ± 0.3
Random-K (SamPO)	51.9 ± 0.3
IAPO	$\mathbf{54.0 \pm 0.4}$

The mean accuracy of IAPO is 4.4 points above DPO and 2.1 points above Random-K. This ordering agrees with the separate single-run comparison in Table 4, supporting the central observation that entropy selection improves on random selection at the same token budget. The repeat experiment covers only these three methods on this backbone: it does not quantify training variation for every model, baseline, or individual benchmark, and the aggregate summaries alone do not establish per-benchmark statistical significance.

6.6 Response Length and Computational Scope

The retained-token budget controls the training loss; generated response length remains an empirical outcome. For Llama3-8B-Base, IAPO produces shorter responses than SamPO on both Arena-Hard (332 versus 422 tokens) and AlpacaEval2 (1,309 versus 1,498), while improving judged quality. For Llama3-8B-Instruct, IAPO produces longer responses than SamPO on both evaluations, and R-DPO yields shorter responses than IAPO across both Llama3 settings. Together with the length-controlled win rates, these results support improved quality without implying that IAPO universally minimizes output length.

Computationally, both policy and reference still score the complete response sequences. Entropy ranking uses the policy's vocabulary distribution at each valid response position, and the resulting discrete masks select terms for the pairwise loss. Consequently, reducing the number of loss terms does not reduce the context length or remove the full-prefix forward computation. The additional entropy and selection operations introduce overhead, whose size depends on the implementation and execution setting. The comparisons above establish benchmark performance and the effect of token selection; they do not establish a training-speed or memory advantage.

7 Conclusion and Limitations

We introduced IAPO, which selects high-entropy positions under an equal retained-token budget for pairwise preference optimization. Matched-budget ablations support the contribution of entropy ranking beyond length equalization, and three independent training runs support this ordering on Qwen2.5-Math-7B. Across the evaluated model settings, IAPO achieves strong average scores, while individual benchmark winners vary.

Limitations. Model entropy is a heuristic ranking signal, not a token-level preference attribution. Our analysis establishes true-distribution information bounds for fixed subsets and optimality for an entropy surrogate; it does not characterize the optimal policy or convergence of adaptive masking. The method adds entropy-computation overhead, and the independent training repeats cover only the central Qwen2.5-Math-7B comparison. Open-ended alignment evaluation is limited to Llama3, and we do not directly compare against TDPO, β-DPO, or BPO. Extending model coverage and testing richer selection signals are useful directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search and summarization, manuscript drafting, organization and language editing, figure and table formatting, review of theoretical exposition, discussions of experimental methodology and result interpretation, and code correction and testing. Generative models were also used for response generation and automated evaluation, as described in Appendix D. The authors take responsibility for the final manuscript, reported results, and conclusions.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

AI@Meta. Llama 3 model card, 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pp. 1352–1361. PMLR, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Yeongmin Kim, HeeSun Bae, Byeonghu Na, and Il-chul Moon. Preference optimization by estimating the ratio of the data distribution. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1987.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. `https://huggingface.co/datasets/AI-MO/NuminaMath-CoT`, 2024.

Tianle Li*, Wei-Lin Chiang*, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024. URL `https://lmsys.org/blog/2024-04-19-arena-hard/`.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 5 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL `https://aclanthology.org/2022.acl-long.229`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. *arXiv preprint arXiv:2406.10957*, 2024.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, August 2024.

Xuan Qi, Jiahao Qiu, Xinzhe Juan, Yue Wu, and Mengdi Wang. Shallow preference signals: Large language model aligns even better with truncated data? *arXiv preprint arXiv:2505.17122*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Soumya Rani Samineni, Durgesh Kalwar, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. RL in name only? analyzing the structural assumptions in RL post-training for LLMs. In *First Workshop on Foundations of Reasoning in Language Models*, 2025. URL `https://openreview.net/forum?id=4vNW1MvX2m`.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2859–2873, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Muning Wen, Junwei Liao, Cheng Deng, Jun Wang, Weinan Zhang, and Ying Wen. Entropy-regularized token-level policy optimization for language agent reinforcement. *arXiv preprint arXiv:2402.06700*, 2024.

Lilian Weng. Reward hacking in reinforcement learning. *lilianweng.github.io*, Nov 2024. URL https://lilianweng.github.io/posts/2024-11-28-reward-hacking/.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-DPO: Direct preference optimization with dynamic β. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4128.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58348–58365. PMLR, 2024. URL https://proceedings.mlr.press/v235/zeng24c.html.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A PROOFS

A.1 PROOF OF PROPOSITION 4.1

Proof. For every token position t, the definition of conditional mutual information gives

$$I_P(Y_t; Z \mid X, Y_{<t}) = H_P(Y_t \mid X, Y_{<t}) - H_P(Y_t \mid X, Y_{<t}, Z).$$

Since conditional entropy is nonnegative,

$$H_P(Y_t \mid X, Y_{<t}, Z) \geq 0,$$

which immediately implies

$$I_P(Y_t; Z \mid X, Y_{<t}) \leq H_P(Y_t \mid X, Y_{<t}).$$

Conditioning on a fixed prefix $(X = x, Y_{<t} = y_{<t})$ yields the pointwise form

$$I_P(Y_t; Z \mid X = x, Y_{<t} = y_{<t}) \leq h_t^\star(x, y_{<t}).$$

□

A.2 PROOF OF PROPOSITION 4.2

Proof. Let $S^c = \{t_1, \ldots, t_m\}$ with $t_1 < \cdots < t_m$. By the chain rule for conditional mutual information,

$$I_P(Y_{S^c}; Z \mid X, Y_S) = \sum_{j=1}^{m} I_P\big(Y_{t_j}; Z \mid X, Y_S, Y_{t_1}, \ldots, Y_{t_{j-1}}\big).$$

For each j, define

$$C_j := (X, Y_S, Y_{t_1}, \ldots, Y_{t_{j-1}}).$$

Then

$$I_P(Y_{t_j}; Z \mid C_j) = H_P(Y_{t_j} \mid C_j) - H_P(Y_{t_j} \mid C_j, Z) \leq H_P(Y_{t_j} \mid C_j).$$

Because S is deterministic and $t_1 < \cdots < t_m$, the conditioning set C_j contains all tokens with index smaller than t_j: every index $r < t_j$ lies either in S or among $t_1, \ldots, t_{j-1}$. Hence $Y_{<t_j}$ is measurable with respect to C_j, and conditioning reduces entropy, so

$$H_P(Y_{t_j} \mid C_j) \leq H_P(Y_{t_j} \mid X, Y_{<t_j}).$$

Summing over j gives

$$I_P(Y_{S^c}; Z \mid X, Y_S) \leq \sum_{j=1}^{m} H_P(Y_{t_j} \mid X, Y_{<t_j}) = \sum_{t \in S^c} H_P(Y_t \mid X, Y_{<t}).$$

□

Corollary A.1 (Posterior closeness induced by small discarded entropy)**.** *For any deterministic subset $S \subseteq [T]$,*

$$\mathbb{E}_P\Big[\|P(Z \mid X, Y) - P(Z \mid X, Y_S)\|_{\mathrm{TV}}^2\Big] \leq \frac{1}{2} \sum_{t \in S^c} H_P(Y_t \mid X, Y_{<t}).$$

A.3 PROOF OF COROLLARY A.1

Proof. The conditional mutual information identity yields

$$I_P(Y_{S^c}; Z \mid X, Y_S) = \mathbb{E}_P[\mathrm{KL}(P(Z \mid X, Y) \,\|\, P(Z \mid X, Y_S))]\,.$$

Applying Pinsker's inequality pointwise inside the expectation gives

$$\|P(Z \mid X, Y) - P(Z \mid X, Y_S)\|_{\mathrm{TV}}^2 \leq \frac{1}{2}\mathrm{KL}(P(Z \mid X, Y) \,\|\, P(Z \mid X, Y_S))\,.$$

Taking expectations on both sides yields

$$\mathbb{E}_P\Big[\|P(Z \mid X, Y) - P(Z \mid X, Y_S)\|_{\mathrm{TV}}^2\Big] \leq \frac{1}{2} I_P(Y_{S^c}; Z \mid X, Y_S).$$

The final bound follows by substituting Proposition 4.2. □

A.4 Proof of Proposition 4.3

Proof. For any subset $S \subseteq [|y|]$ with $|S| = K$,

$$U_x^\star(y, S) = \sum_{t \notin S} h_t^\star(x, y_{<t}) = \sum_{t=1}^{|y|} h_t^\star(x, y_{<t}) - \sum_{t \in S} h_t^\star(x, y_{<t}).$$

The first term on the right-hand side is constant over all feasible subsets S of size K. Therefore minimizing $U_x^\star(y, S)$ is equivalent to maximizing the retained entropy mass $\sum_{t \in S} h_t^\star(x, y_{<t})$. Any maximizer $S_K^\star(x, y)$ of the latter is therefore also a minimizer of the former. □

Corollary A.2 (Pairwise equal-budget selection)**.** *Consider a preference pair (x, y^+, y^-) and define the common retained-token budget*

$$K := \min(|y^+|, |y^-|).$$

For each response $y \in \{y^+, y^-\}$, let $S(y) \subseteq [|y|]$ satisfy $|S(y)| = K$. Then the optimization problem

$$\min_{\substack{|S(y^+)|=K \\ |S(y^-)|=K}} U_x^\star(y^+, S(y^+)) + U_x^\star(y^-, S(y^-))$$

is minimized by selecting the top-K entropy positions in each response. In particular, the shorter response is kept unchanged.

A.5 Proof of Corollary A.2

Proof. The objective

$$U_x^\star(y^+, S(y^+)) + U_x^\star(y^-, S(y^-))$$

separates additively across the two responses, and the constraints $|S(y^+)| = K$ and $|S(y^-)| = K$ are independent. Therefore the joint minimization decomposes into two independent fixed-budget problems, one for y^+ and one for y^-. By Proposition 4.3, each subproblem is solved by retaining the top-K entropy positions of the corresponding response.

If y^{short} is the shorter response, then by definition $K = |y^{\text{short}}|$. The only subset of $[|y^{\text{short}}|]$ with cardinality K is the full set itself, so

$$S(y^{\text{short}}) = [|y^{\text{short}}|].$$

Hence the shorter response is left unchanged. □

Proposition A.3 (Near-optimality under entropy estimation error)**.** *Fix a prompt-response pair (x, y) and a token budget $K \le |y|$. Let*

$$S_K^\star(x, y) \in \arg\max_{\substack{S \subseteq [|y|] \\ |S|=K}} \sum_{t \in S} h_t^\star(x, y_{<t}), \qquad \widehat{S}_K(x, y) \in \arg\max_{\substack{S \subseteq [|y|] \\ |S|=K}} \sum_{t \in S} \widehat{h}_t(x, y_{<t}).$$

Suppose that for all $t \in [|y|]$,

$$\left| \widehat{h}_t(x, y_{<t}) - h_t^\star(x, y_{<t}) \right| \le \delta.$$

Then

$$U_x^\star(y, \widehat{S}_K(x, y)) \le U_x^\star(y, S_K^\star(x, y)) + 2K\delta.$$

A.6 Proof of Proposition A.3

Proof. Define

$$f(S) := \sum_{t \in S} h_t^\star(x, y_{<t}), \qquad g(S) := \sum_{t \in S} \widehat{h}_t(x, y_{<t}).$$

For every feasible subset S with $|S| = K$, the uniform error assumption implies

$$|f(S) - g(S)| \le \sum_{t \in S} \left| h_t^\star(x, y_{<t}) - \widehat{h}_t(x, y_{<t}) \right| \le K\delta.$$

Because $S_K^\star(x, y)$ maximizes f and $\widehat{S}_K(x, y)$ maximizes g,

$$f(S_K^\star) \le g(S_K^\star) + K\delta \le g(\widehat{S}_K) + K\delta \le f(\widehat{S}_K) + 2K\delta.$$

Rearranging gives

$$f(\widehat{S}_K) \ge f(S_K^\star) - 2K\delta.$$

Now write

$$U_x^\star(y, S) = \sum_{t=1}^{|y|} h_t^\star(x, y_{<t}) - f(S).$$

The first term is constant over all feasible subsets, so the previous inequality implies

$$U_x^\star(y, \widehat{S}_K) \le U_x^\star(y, S_K^\star) + 2K\delta.$$

This proves the stated surrogate bound under the uniform entropy-error assumption. It is not a bound on preference information under an adaptive mask. □



(a) Frequent tokens with high entropy

(b) Frequent tokens with low entropy

Figure 2: **Word cloud visualization of high and low entropy tokens.** Tokens are aggregated by token type, and the occurrence counts and average entropy are computed. The high entropy cloud (left) shows tokens from the upper entropy quantile, while the low entropy cloud (right) shows tokens from the lower quantile. Font size reflects a balanced weight combining normalized entropy and frequency.

B Word Cloud Visualization of Token Entropy

To provide an intuitive view of the non-uniform distribution of token-level information, we visualize tokens with high entropy and low entropy using word clouds constructed from the UltraFeedback (Cui et al., 2023) dataset under the SFT-trained Llama3-8B (AI@Meta, 2024) model.

Tokens are grouped by token type, and for each type, we aggregate its occurrence frequency across the dataset and compute its average entropy. We filter non-linguistic artifacts, special symbols, numeric fragments, and rare tokens, retaining only tokens appearing at least twenty times. High entropy tokens are selected from the upper entropy quantile, while low entropy tokens are selected from the lower quantile of the entropy distribution. To balance semantic salience and statistical prevalence, each token is assigned a combined weight based on normalized mean entropy and normalized frequency, and the top tokens are rendered as word clouds, where font size reflects the resulting balanced weight.

The clouds visualize differences in the token types associated with the upper and lower entropy quantiles. Aggregating by token type does not measure a token's contribution to a preference label, so this visualization is descriptive rather than a test of preference relevance.

C Impact of Length Difference on Preference

Figure 3 shows binned AlpacaEval 2 win rates as a function of the response-length difference. The curves are descriptive and depend on the retained bins and smoothing; they do not establish a causal effect of masking on length sensitivity. The aggregate comparison is reported in Table 1.

D Experimental Details

In this section, we provide details of the training configurations and dataset preparation to support reproducibility and further validate the effectiveness of IAPO. Code is available at `https://anonymous.4open.science/r/IAPO`.

D.1 Training Details

Setup. We implement IAPO using the TRL repository to ensure a fair comparison. Baseline objectives are summarized in Table 9; the corresponding reference implementations are SimPO `https://github.com/princeton-nlp/SimPO` and SamPO `https://github.com/LuJunru/SamPO`. Model-level training settings are reported in Table 7; its β row refers to DPO. For the Llama3-8B-Instruct comparison, DPO uses $\beta = 0.01$, while SimPO uses $\beta = 2.5$ and $\gamma/\beta = 0.55$. Both

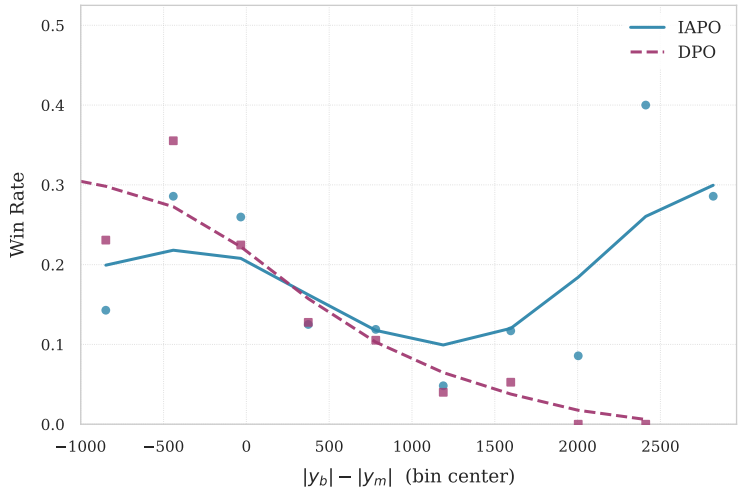


Figure 3: **Win Rate vs Length Difference.** For each AlpacaEval 2 sample, we compute the response length difference $|y_b| - |y_m|$ between the GPT-4 Turbo baseline output and the evaluated model output; positive values indicate that the baseline response is longer. We then partition the observed range into equal-width bins and report win rates within each remaining bin after dropping bins with fewer than three samples. Curves are obtained by applying a one-dimensional Gaussian smoothing filter ($\sigma = 1.2$) to the retained bin means for visualization only.

Table 6: Prompts for Mathematical Reasoning.

Prompts for Mathematical Reasoning
<|im_start|>system\n Please reason step by step, and put your final answer within \boxed{}.<|im_end|>\n <|im_start|>user\n Question Let's think step by step and output the final answer within \boxed<|im_end|>\n <|im_start|>assistant

use the `princeton-nlp/llama3-ultrafeedback` dataset, learning rate 10^{-6}, one epoch, global batch size 128, maximum prompt/sequence lengths 1,800/2,048, 10% warmup, AdamW, cosine decay, and seed 42. The independent-training comparison in Section 6.5 is reported separately from the single-run benchmark tables. All experiments are conducted on 8×NVIDIA A800 GPUs with Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz.

Dataset. For *textual understanding tasks,* the dataset preparation is aligned with SimPO (Meng et al., 2024). For *mathematical reasoning tasks,* we sample 20K prompts from the NuminaMath (LI et al., 2024) dataset. For each prompt, 8 responses are generated using the chain-of-thought prompt template shown in Tab. 6 and ranked by a rule-based reward. Responses with the highest and lowest reward are selected as the chosen and rejected outputs to form preference pairs, while prompts yielding identical reward across responses are excluded. These final preference pairs are used for training our method and all other baseline methods to ensure a fair comparison.

D.2 Evaluation Details

For *textual understanding tasks,* we provide the summarization of the evaluation benchmarks used in our experiments in Tab. 8. We also provide the prompt templates employed for AlpacaEval 2.0 (Li et al., 2023) in Tab. 10 and for Arena-Hard v0.1 in Tab. 13. For *mathematical reasoning tasks,* we use the prompt in Tab. 6 and evaluate all methods with zero-shot chain-of-thought prompting under greedy decoding.

Table 7: Hyperparameters for different models.

Setting \ Model	Llama3-8B-Base	Llama3-8B-Instruct	Gemma2-9B-it	Qwen2.5-Math-1.5B	Qwen2.5-Math-7B
Learning rate	1.0e-6		8.0e-7	1.0e-6	
DPO beta β	0.01			2.0	
Max prompt length	1800			1000	
Max length	2048			4096	
Warmup ratio	0.1 (10% steps)			0.0	
Epoch	1			2	
Optimizer	AdamW (Loshchilov & Hutter, 2017)				
Learning rate scheduler	Cosine				
Global batch size	128				
Seed	42				

Table 8: **Evaluation benchmarks for textual understanding tasks.** The baseline model refers to the model being compared against. GPT-4 Turbo corresponds to GPT-4-1106-Preview.

	# Exs.	Baseline Model	Judge Model	Scoring Type	Metric
AlpacaEval 2.0 Li et al. (2023)	805	GPT-4 Turbo	GPT-4 Turbo	Pairwise comparison	LC & raw win rate
Arena-Hard v0.1 Li* et al. (2024)	500	GPT-4-0314	GPT-4 Turbo	Pairwise comparison	Win rate
IFEval Zhou et al. (2023)	541	-	-	Rule-based checking	Prompt/Instruction & Strict/Loose Acc.
HellaSwag Zellers et al. (2019)	10042	-	-	Rule-based checking	Accuracy
TruthfulQA Lin et al. (2022)	817	-	-	Answer likelihoods	MC1 & MC2 scores
PiQA Bisk et al. (2020)	1838	-	-	Rule-based checking	Accuracy

D.3 Baselines

To ensure a fair and transparent comparison, we implement representative preference optimization baselines that explicitly address length bias through modified aggregation or normalization strategies. R-DPO (Park et al., 2024) introduces explicit length regularization, SamPO (Lu et al., 2024) uniformly down-samples tokens to equalize comparison length, and SimPO (Meng et al., 2024) normalizes log probability by response length.

Table 9 summarizes the per-pair objective definitions. Write $\ell_t^y = \log[\pi_\theta(y_t \mid x, y_{<t})/\pi_{\text{ref}}(y_t \mid x, y_{<t})]$ and $q_t^y = \log \pi_\theta(y_t \mid x, y_{<t})$. For SamPO, S^w and S^l are uniform subsets of their respective response positions, each of size $K = \min(|y^w|, |y^l|)$; training additionally averages over this sampling. Each selected position uses the same mask in its policy/reference log-ratio and retains its full original prefix. The R-DPO length correction belongs inside the logistic logit.

Table 9: Per-pair objectives for DPO and comparison methods. Full response prefixes are included in ℓ_t^y and q_t^y as defined above.

Method	Per-pair loss	Aggregation
DPO (Rafailov et al., 2023)	$-\log \sigma\big(\beta[\sum_t \ell_t^w - \sum_t \ell_t^l]\big)$	All positions
R-DPO (Park et al., 2024)	$-\log \sigma\big(\beta[\sum_t \ell_t^w - \sum_t \ell_t^l] + \alpha[\lvert y^w\rvert - \lvert y^l\rvert]\big)$	Length correction
SamPO (Lu et al., 2024)	$-\log \sigma\big(\beta[\sum_{t\in S^w} \ell_t^w - \sum_{t\in S^l} \ell_t^l]\big)$	Random equal-budget subsets
SimPO (Meng et al., 2024)	$-\log \sigma\Big(\frac{\beta}{\lvert y^w\rvert}\sum_t q_t^w - \frac{\beta}{\lvert y^l\rvert}\sum_t q_t^l - \gamma\Big)$	Average log-probability
IAPO	$-\log \sigma\big(\beta[\sum_t m_t^w \ell_t^w - \sum_t m_t^l \ell_t^l]\big)$	Entropy-ranked equal budget

E Detailed Token Selection Ablation Results

This section provides per-benchmark breakdowns for the ablation study summarized in Section 6.4. We compare full-sequence aggregation (DPO), random-K selection (R-K, equivalent to SamPO (Lu et al., 2024)), bottom-K entropy selection (B-K), two one-sided masking variants, and the full IAPO objective. Tab. 11 reports results on textual alignment benchmarks (Llama3-8B-Base) and Tab. 12 on mathematical reasoning benchmarks (Qwen2.5-Math-7B). The per-benchmark breakdowns support the main-text summary: bottom-K is weaker on the aggregate metrics, random-K improves over DPO on those metrics, and entropy-guided selection improves further. The one-sided variants are competitive on several individual metrics, with Rej-only strongest on AlpacaEval2 raw win rate and

Table 10: Prompts for AlpacaEval 2.0.

Prompts for AlpacaEval 2.0

<|im_start|>system
You are a highly efficient assistant, who evaluates and selects the best large language model (LLMs) based on the quality of their responses to a given instruction. This process will be used to create a leaderboard reflecting the most accurate and human-preferred answers.
<|im_end|>
<|im_start|>user
I require a leaderboard for various large language models. I'll provide you with prompts given to these models and their corresponding outputs. Your task is to assess these responses, and select the model that produces the best output from a human perspective.

Instruction

```
{
   "instruction": """instruction""",
}
```

Model Outputs

Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.

```
{
   {
      "model_identifier": "m",
      "output": """output_1"""
   },
   {
      "model_identifier": "M",
      "output": """output_2"""
   }
}
```

Task

Evaluate the models based on the quality and relevance of their outputs, and select the model that generated the best output. Answer by providing the model identifier of the best model. We will use your output as the name of the best model, so make sure your output only contains one of the following model identifiers and nothing else (no quotes, no spaces, no new lines, ...): m or M.

Best Model Identifier
<|im_end|>

Cho-only strongest on Minerva Math and AMC2023, but the full IAPO objective remains best on the reported averages.

E.1 Independent Training Runs

The independent-training results and their aggregation protocol are reported in Section 6.5 and Table 5. Each run contributes one equally weighted eight-benchmark average, and the reported standard deviation is taken across three trained models per method. The detailed single-run scores in Tables 3 and 12 are separate from these repeat summaries and should not be interpreted as the per-benchmark means of the three runs.

Table 11: **Token selection and masking ablation.** We compare full-sequence aggregation (DPO), random-K (R-K, SamPO), bottom-K (B-K), one-sided masking variants, and the full top-K entropy selection objective. Results are reported on Arena-Hard and AlpacaEval 2 using the Llama3-8B-Base model; the 95% CI column reports the lower and upper deviations around the Arena-Hard WR estimate.

Method	Arena-Hard			AlpacaEval2			
	WR(%)	95% CI	Avg. #tokens	WR(%)	LC(%)	Std Error	Avg. #tokens
SFT	3.3	$(-0.6, 0.8)$	389	3.7	5.7	0.66	1000
DPO	15.9	$(-1.6, 1.6)$	542	13.3	15.5	1.20	1708
R-K	25.4	$(-2.0, 1.1)$	422	14.25	18.74	1.11	1498
B-K	24.9	$(-1.8, 2.0)$	430	7.62	8.66	1.22	1543
Rej-only	25.3	$(-0.9, 1.1)$	477	**15.52**	17.61	1.41	1622
Cho-only	<u>26.1</u>	$(-1.3, 1.5)$	399	14.25	<u>18.77</u>	0.98	1451
IAPO	**27.3**	$(-1.8, 1.5)$	332	<u>14.42</u>	**20.18**	1.13	1309

Table 12: **Token selection and masking ablation.** We compare full-sequence aggregation (DPO), random-K (R-K, SamPO), bottom-K (B-K), one-sided masking variants, and the full top-K entropy selection objective. Results are reported on mathematical reasoning tasks using Qwen2.5-Math-7B. Average is the equally weighted mean of the eight displayed scores, rounded to one decimal place.

Method	GSM8K	MATH	Minerva Math	Olympiad Bench	AIME 2024	AMC 2023	GaoKao 2023 En	College Math	Average
Baseline	63.5	65.4	9.2	22.5	23.3	52.5	44.9	31.2	39.1
DPO	80.0	74.8	21.3	**38.1**	<u>26.7</u>	52.5	57.1	44.3	49.4
R-K	84.0	<u>76.0</u>	24.3	36.9	23.3	**62.5**	60.0	<u>45.3</u>	51.5
B-K	66.3	69.2	12.5	25.3	<u>26.7</u>	47.5	46.5	35.7	41.2
Rej-only	85.4	75.2	<u>33.5</u>	37.3	**30.0**	57.5	60.5	40.3	52.5
Cho-only	<u>85.6</u>	**76.4**	**36.0**	37.0	<u>26.7</u>	**62.5**	<u>61.8</u>	40.0	<u>53.3</u>
IAPO	**86.4**	**76.4**	<u>33.5</u>	<u>37.8</u>	**30.0**	<u>60.0</u>	**63.6**	**45.4**	**54.1**

Table 13: Prompts for ArenaHard v0.1.

Prompts for ArenaHard v0.1

system_prompt:
"Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.\n\n
Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.\n\n
When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.\n\n
Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.\n\n
Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.\n\n
After providing your explanation, you must output only one of the following choices as your final verdict with a label:\n\n
1. Assistant A is significantly better: [[A>>B]]\n
2. Assistant A is slightly better: [[A>B]]\n
3. Tie, relatively the same: [[A=B]]\n
4. Assistant B is slightly better: [[B>A]]\n
5. Assistant B is significantly better: [[B>>A]]\n\n
Example output: "My final verdict is tie: [[A=B]]"."

prompt_template:
"<|User Prompt|>\n
QUESTION\n\n
<|The Start of Assistant A's Answer|>\n
ANSWER_A\n
<|The End of Assistant A's Answer|>\n\n
<|The Start of Assistant B's Answer|>\n
ANSWER_B\n
<|The End of Assistant B's Answer|>"